Sleep, 19(1):26–35

Automatic Analysis and Sleep Scoring

Sleep Stage Scoring Using the Neural Network Model: Comparison Between Visual and Automatic Analysis in Normal Subjects and Patients

*N. Schaltenbrand, †R. Lengelle, *M. Toussaint, *R. Luthringer,
*G. Carelli, *A. Jacqmin, *E. Lainey, ‡A. Muzet and *J. P. Macher

**Institute for Research in Neurosciences and Psychiatry, Rouffach;*
†University of Troyes, Troyes; and ‡Laboratoire de Physiologie et de Psychologie Environmentale, CNRS, Strasbourg, France

Summary: In this paper, we compare and analyze the results from automatic analysis and visual scoring of nocturnal sleep recordings. The validation is based on a sleep recording set of 60 subjects (33 males and 27 females), consisting of three groups: 20 normal control subjects, 20 depressed patients and 20 insomniac patients treated with a benzodiazepine. The inter-expert variability estimated from these 60 recordings (61,949 epochs) indicated an average agreement rate of 87.5% between two experts on the basis of 30-second epochs. The automatic scoring system, compared in the same way with one expert, achieved an average agreement rate of 82.3%, without expert supervision. By adding expert supervision for ambiguous and unknown epochs, detected by computation of an uncertainty index and unknown rejection, the automatic/expert agreement grew from 82.3% to 90%, with supervision over only 20% of the night. Bearing in mind the composition and the size of the test sample, the automated sleep staging system achieved a satisfactory performance level and may be considered a useful alternative to visual sleep stage scoring for large-scale investigations of human sleep. **Key Words:** Sleep—Automatic scoring—Visual scoring—Scoring variability—Neural networks.

Several automatic sleep scoring systems have been described in the last 2 decades. First, hybrid systems were used (1–3). Second, in addition to these heuristic approaches, a considerable number of methods based on statistical pattern recognition techniques, which utilize more formal approaches, were devised (4–8). Third, expert systems were also designed (9,10) and neural networks were used (11,12). Reliability of automatic sleep scoring systems has been reported. Smith et al. (3) presented an agreement of 83% between human scoring and their hybrid system of staging, but they pooled stage 1 and rapid eye movement (REM) sleep. Gaillard and Tissot (2) found an agreement of 77.8% using the same data as Smith and Karacan (3) and 82.7% with their own data (stage 1 and REM also pooled). Martin et al. (4) found 82% agreement with nine healthy young subjects, and Hasan (13) reported 80% agreement in young normal subjects, 77% for older normal subjects and 75% for alcoholics (using 20-second epochs). Stanus et al. (7) found 75% agreement on 15 control subjects and 70% on 15 patients, whereas Kuwahara et al. (8) found 89.1% agreement in 12 control subjects. Automatic analysis with the Oxford sleep stager showed 73.1% agreement in 10 subjects without sleep disturbances (14).

Several studies have also evaluated agreement in human vs. human scoring. Agreement between scorers ranged from 88% to 85% according to Gaillard and Tissot (2) and Smith et al. (3), on the basis of 60-second epochs. Recently Stanus et al. (7) obtained an 82% reliability using 20-second epoch definition, and Ferri et al. (15) obtained 80% agreement with nine groups of readers from different laboratories. Kubicki et al. (14) found an agreement rate of 91.3% between two independent readers. Agreement in visual scoring of sleep stages among 10 laboratories in Japan (16) showed values ranging from 67.3% to 75.3% for two healthy subjects. Inter-reader agreement between dif-

Accepted for publication October 1995.

Address correspondence and reprint requests to Nicolas Schaltenbrand, Ph.D., Institute for Research in Neurosciences and Psychiatry, Centre Hospitalier, Sector 8, 68250 Rouffach, France.

ferent laboratories is broadly similar. Variations are based on different time bases and on the number of readers for the validation. The difference between readers has three major causes: 1) the scoring of slow-wave sleep, which illustrates the human subjectivity introduced into the rules applied to the threshold of 75 μV for delta wave recognition and in the detection of the percentage of delta waves in the epochs; 2) the scoring of stage 1 sleep, which is subject to great variation due to a lack of clearly characteristic features; and 3) application of "3-minute rules" in the standardized system (17).

Our group is working on signal processing techniques and pattern recognition applied to biomedical systems. Our main goal is to define a structure and an environment for electroencephalographic (EEG) signal interpretation in medicine. In this field, a sleep analysis system was developed. We have applied a new method for detecting sleep stage patterns in data that is based on a neural network model. The originality of this method lies in *knowledge learning with a small structure,* which allows real-time classification and handling of the large diversity of sleep data. Our ultimate objective is the use of the automatic sleep system in an everyday environment. Therefore, we paid special attention to validation of the results against the standardized scoring technique using data collected from several populations.

METHODS

Subjects

Sleep recordings from 60 subjects (33 males and 27 females) were analyzed. The subjects included 20 normal controls (12 males and 8 females), aged 19 through 42 years (mean = 26 ± 7); 20 patients with depression (10 males and 10 females), aged 35 through 65 years (mean = 48 ± 10); and 20 patients with insomnia treated with benzodiazepines (11 males and 9 females), aged 25 through 72 years (mean = 44 ± 11).

The normal control subjects were volunteers screened to rule out those with significant medical problems, major sleep disorders or abnormal sleep habits.

Patients in the group with depression were diagnosed according to DSM-III-R (18) criteria and had a minimum total score of 17 on the Hamilton Depression Scale (19). All patients were drug-free for at least 5 weeks at the time of study.

Patients with insomnia were diagnosed on the basis of a medical history, full clinical examination and clinical laboratory tests. All met DSM-III-R criteria for primary insomnia. No patients in the insomnia group had sleep difficulties that were obviously secondary to psychopathology (e.g. major affective disorders, psychosis) or medical problems (e.g. pain). All the insomniac patients had been treated with a benzodiazepine for more than 6 months. This group was examined in order to test our system not only in "good sleepers" but also in patients who might exhibit large variations in their EEG patterns. Thus, typical EEG modifications due to benzodiazepine treatment might be expected, such as increase in beta and spindle activity or decrease in slow-wave activity (20).

The three groups were chosen to provide samples of normal and pathological sleep recordings. Testing only one set of recordings seemed inadequate. In this study, we focused our work on three major groups that together characterize the diversity of the different types of sleep recordings made in our laboratory.

Recordings

Polygraphic sleep recordings used for this evaluation did not include adaptation night data. Data channels included two EEG channels (C4-A1 and 01-C3), one horizontal electrooculogram (EOG) channel and one chin electromyogram (EMG) channel.

An electronic device was used for amplification and analogue filtering (type: Bessel order 2) with the following cut-off frequencies: 0.5–30 Hz for EEG, 0.5–15 Hz for EOG and 5–70 Hz for EMG. The signals were digitized at a sampling rate of 128 Hz.

Automatic sleep scoring system

Real time computerized analysis of the sleep recordings was performed in three major steps: 1) Feature extraction from three data channels (central EEG, EOG, EMG) using signal processing techniques. The result, for each 30 second epoch, is a "feature vector" of 17 components. 2) Classification of each "feature vector" characterizing each page to one of the sleep stages using a multilayer neural network model. 3) Supervision of the classification using an uncertainty index computed in the output space of the multilayer network and distance rejection using a nonsupervised neural network called ART2.

Feature extraction

The automatic sleep stage scoring system requires only three data channels: central EEG, horizontal EOG and chin EMG. This limitation was chosen to comply with the minimum recommendation of the standardized system and to allow the implementation of the system with a portable device. The feature extraction technique (Fig. 1) involved segmenting the signal into

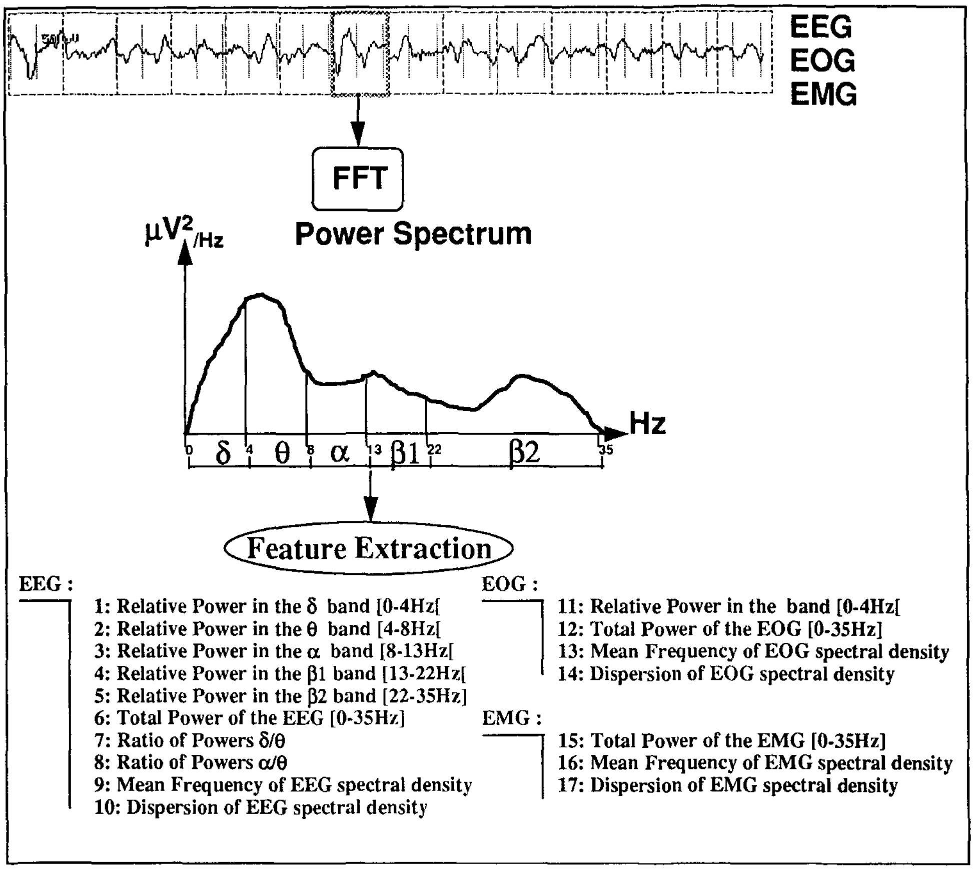
EEG
EOG
EMG
FFT
µV²/Hz
Power Spectrum
Hz
0 δ 4 θ 8 α 13 β1 22 β2 35
Feature Extraction
EEG :
1: Relative Power in the δ band [0-4Hz[
2: Relative Power in the θ band [4-8Hz[
3: Relative Power in the α band [8-13Hz[
4: Relative Power in the β1 band [13-22Hz[
5: Relative Power in the β2 band [22-35Hz]
6: Total Power of the EEG [0-35Hz]
7: Ratio of Powers δ/θ
8: Ratio of Powers α/θ
9: Mean Frequency of EEG spectral density
10: Dispersion of EEG spectral density
EOG :
11: Relative Power in the band [0-4Hz[
12: Total Power of the EOG [0-35Hz]
13: Mean Frequency of EOG spectral density
14: Dispersion of EOG spectral density
EMG :
15: Total Power of the EMG [0-35Hz]
16: Mean Frequency of EMG spectral density
17: Dispersion of EMG spectral density

FIG. 1. Feature extraction technique.

short fixed nonoverlapping intervals of 2 seconds. The power spectrum was calculated using fast Fourier transformation with a 2-second epoch (256 points). Truncating error was reduced with a Hanning window. Characteristics of the power spectrum called "features" were then extracted. Feature extraction involves the reduction of large amounts of data to a meaningful summary. Fourier transformation is useful in problems where the amplitude or energy spectrum exhibits significant interclass differences. No theory exists to determine which particular transform will be most beneficial. Often transform applicability is best evaluated by measuring recognition system performance using transformed data. Techniques for defining features that carry significant information may be divided into human or logical design techniques (known features) and automatic design techniques (statistical features). Our first step was to define features using our considerable experience in manual sleep stage classification. We divided the spectrum into EEG frequency bands (delta, theta, alpha, beta1 and beta2). Next, we used statistical tests (*t* test, analysis of variance) or direct examination of the transformed data (one-dimensional plot or principal component analysis) to validate parameters. Mean spectrum frequency and dispersion were added for discriminant power in univariate statistical tests. Owing to the nonlinear classifier used, it is difficult to determine the exact importance of each individual parameter. The final performance of the recognition system will reflect the parameter quality. Seventeen time-frequency parameters are estimated for each 30-second epoch by averaging 15 successive 2-second estimates.

Classification

The aim of the classification to assign each vector of parameters to one of the sleep stages. We focused our research on supervised classification algorithms that could take into account the huge variability of data. We used a neural network model for the classification of sleep stages; its properties have been published previously (12,21). The neural network model used for the sleep stage classification is a multilayer perceptron. The model we used for this study is shown in Fig. 2. Here the input value X_k is the parameter vector previously described. The units on the input layer are connected to the units on the hidden layer that in turn are connected to the units on the output layer. The goal of the network is to correctly classify the 30-second epochs of the sleep recording, charac-

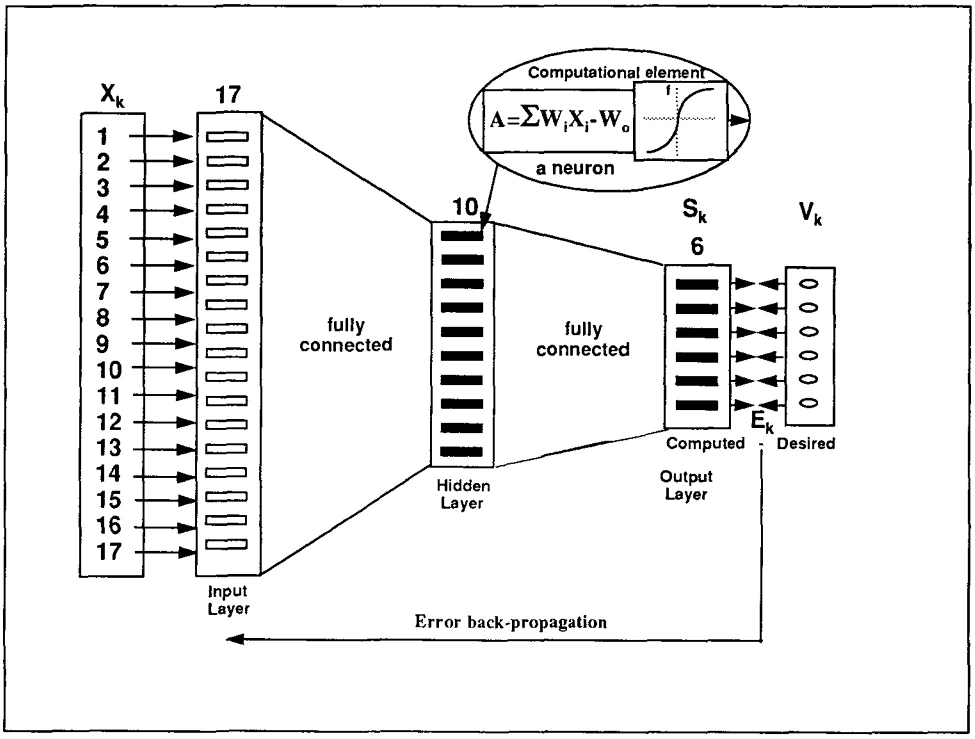

FIG. 2. Neural network model for sleep stage classification.

terized by the 17 time-frequencies parameter vector, to one of the different sleep stages. Thus the output layer is made of six units, one for each stage (or class). The first step in neural networks is the learning phase. Learning consists of altering the values of weights in response to a teaching signal that provides information about the correct classification of the input patterns. The learning algorithms are based on the minimization of a "cost" function. Adaptation or learning is a major focus of neural network research. We used the well-known "back-propagation algorithm" (22). The learning set used in this study consisted of 12 all-night sleep recordings totally independent from the testing set. We only kept the proportion of the test set, i.e. four normal control subjects, four depressed patients and four insomniac patients. When learning is completed, the main properties of the learning set are concentrated in the connection weights. We used the network for the classification of new independent all-night sleep recordings. This test phase consists of the presentation of the parameter vector for each page at the input layer of the network and computing the forward propagation to obtain the sleep stage at the output layer. Test computing time for 1 night of 1,000 epochs is approximately 5 seconds, which allows real time classification. Moreover, connection weights of the neural net took only 984 bytes of memory.

Supervision

When applying multilayer neural networks to a classification task, one generally uses a learning set as prior information. In diagnostic problems, not all states of the system are usually available. Consequently, the learning set is not exhaustive. In some cases, it is very difficult to indicate the correct scoring. These difficulties have been discussed by Kubicki et al. (23). Thus, some information is needed about the degree of certainty of the decision. In other words, it would be very interesting to know if the decision taken by the automatic system was made without any doubt. Taking into account these considerations we propose two solutions: 1) the *unknown*, characterized by a state for which initial rules are totally inappropriate; and 2) the *uncertainty*. This notion means that, for some inputs, a unique decision may not be reliable.

Rejection of the unknown

The classification method must be able to eliminate artifacts. This can be accomplished by distance rejection, a property of many statistical pattern recognition algorithms. An input pattern can be very far away from any training sample and nevertheless may activate the output neuron corresponding to a predefined class. This is because the decision functions are only valid in a neighborhood of the learning set. One solution is with the addition of an unsupervised network with two functions: first, to approximate the learning set hull; and second, to form clusters of rejected points. The nonsupervised network we used for this task is a very simplified version of ART2, proposed by Carpenter and Grossberg (24), called NeoART (25). NeoART compressed data sets by the automatic extraction of prototypes. At the end of the learning process, the learning set is represented by prototypes. A prototype fires if an observation falls into a hypersphere centered on it and with a radius related to a distance parameter. The choice of the distance parameter must be determined using descriptive multidimensional analysis, according to a compromise between precision and number of prototypes (here we used a distance parameter equal to 0.2). Therefore, the learning set is described by a union of hyperspheres. In order to obtain distance rejection, when learning was completed, we just determined whether at least one output NeoART cell fired. If so, the observation to be classified belonged to a region where the multilayer network output was reliable. Otherwise, a new prototype would be generated by NeoART, and the observation would be rejected. Note that the created prototype can be used to form a cluster with rejected points, if these points fall into its influence region.

Uncertainty rejection

Another important characteristic is uncertainty rejection. For some inputs, a unique decision may not be reliable. To handle this problem, using multilayer networks, we determined whether an input vector belonged to an ambiguous zone in the representation space by evaluating the minimum distance between the network output and the closest theoretical decision. The theoretical decisions Di, $i = 1 \ldots m$ are defined as m-dimensional binary vectors with i^{th} component equal to 1 if the presented observation belongs to cluster i, and -1 elsewhere. The actual output O is continuous in the hypercube $[-1, 1]^m$. For uncertainty rejection, the proposed criterion is:

if $d^* = \min_i$ [distance (O,Di)] $> t$ the observation O is considered as ambiguous.
else the observation O belongs to cluster j.
d^* will be called the uncertainty index (26).

The distance used in all experiments was the Euclidean distance. When a distance exceeded threshold, the input was classified as ambiguous, and the system gave a probability of belonging to the nearest classes. The threshold was determined by minimizing the error probability of the detector by plotting the absolute frequency histograms of correctly classified data and misclassified data versus the distance. Optimum threshold was defined by the abscissa of the intersection between these histograms (i.e. minimum error abscissa = 1.7).

Figure 3 shows the uncertainty index variations during a complete night. The uncertainty index is very high during slow-wave sleep (stage 3 and 4), characterizing the artificial distinction introduced by the delta wave rules.

Sleep scoring validation method

To assess variability between visual and automatic scoring, we compared the 60 nights epoch by epoch. First, inter-scorer variability was performed to define a reference by comparing data epoch by epoch. Then, sleep stage inter-expert agreement was computed. Second, sleep parameter comparisons were made to determine if variability introduced by a scorer influenced final diagnosis. Afterwards, analogous comparisons were used to estimate the agreement between the automatic sleep staging system and the scorers.

RESULTS

Inter-expert agreement

For the three groups of 20 recordings, the two visual scorings were compared epoch by epoch to derive an agreement matrix for each group (Table 1). The numbers of epochs correctly classified by the two experts are on the main diagonal in bold type. Other squares contain the numbers of incorrectly scored epochs.

Control group

Table 1A gives the agreement matrix derived from the 20 control recordings. For example, row S3 indicates that, taking expert 1 as reference, 912 epochs of stage 3 sleep were correctly classified by expert 2, 190 epochs were scored by expert 2 in stage 2 and 171 epochs were scored in stage 4. The TOT column indicates the total number of epochs in each stage for expert 1. The TOT row indicates the total number of epochs in each stage for expert 2. The AGR.% row indicates the agreement percentage for one stage between experts, with expert 2 as reference (example: S3 agreement % = 912/1,495 × 100 = 61%). The AGR.% column indicates the agreement percentage for one stage between experts, with expert 1 as reference (example: S3 agreement % = 912/1,273 × 100 = 71.6%). The last two columns indicate the number and the percentage of difference per stage.

The total agreement percentage over 21,138 epochs that were visually analyzed was 88.2% between the two experts, i.e. 2,502 epochs were scored differently.

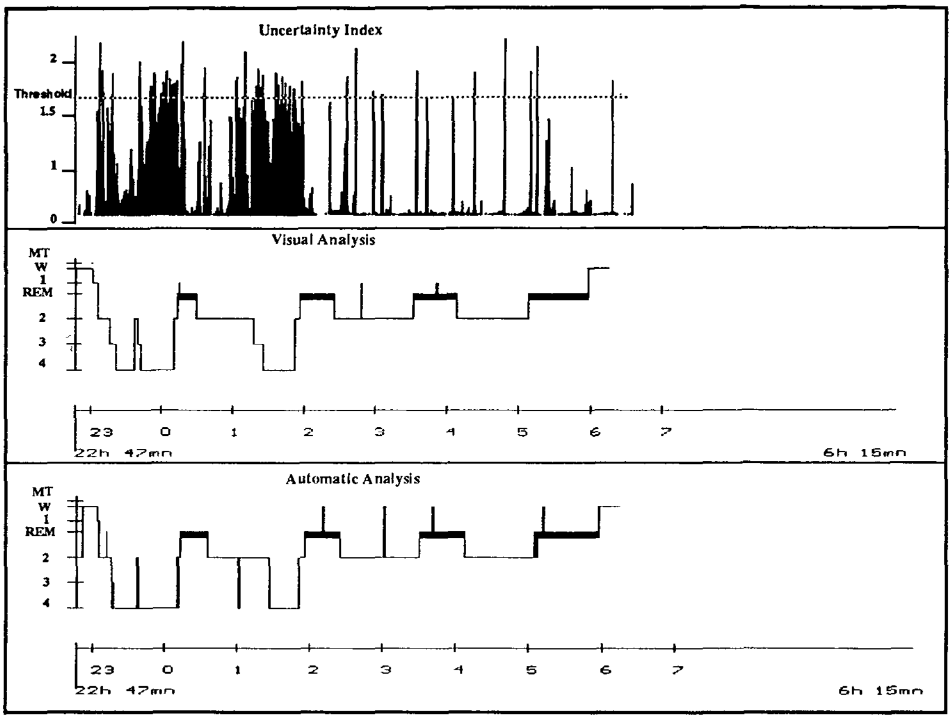

FIG. 3. Evolution of the uncertainty index for a whole night.

TABLE 1. *Agreement matrix inter-experts for the three groups*

		Expert 2									
A. Control		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
Expert 1	W	**1,504**	174	71		1	11	1,761	85.4	257	14.6
	S1	71	**436**	113			29	649	67.2	213	32.8
	S2	54	383	**8,313**	367	8	268	9,393	88.5	1,080	11.5
	S3			190	**912**	171		1,273	71.6	361	28.4
	S4	3		19	216	**2,665**		2.903	91.8	238	8.2
	REM	40	149	164			**4,806**	5,159	93.2	353	6.8
	TOT	1,672	1,142	8,870	1,495	2,845	5,114	**21,138**	**88.2**	2,502	**11.8**
	Agr. %	89.95	38.18	93.72	61.00	93.67	93.98				
		Expert 2									
B. Depressed		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
Expert 1	W	**3,440**	317	74			24	3,855	89.2	415	10.8
	S1	139	**589**	254			54	1,036	56.9	447	43.1
	S2	78	210	**7,628**	847	37	92	8,892	85.8	1,264	14.2
	S3	1		79	**688**	248		1,016	67.7	328	32.3
	S4				61	**754**		815	92.5	61	7.5
	REM	55	127	252			**4,032**	4,466	90.3	434	9.7
	TOT	3,713	1,243	8,287	1,596	1,039	4,202	20,080	**85.3**	2,949	**14.7**
	Agr. %	92.65	47.39	92.05	43.11	72.57	95.95				
		Expert 2									
C. Insomniac		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
Expert 1	W	**2,695**	241	63			14	3,013	89.4	318	10.6
	S1	79	**521**	230		1	34	865	60.2	344	39.8
	S2	67	141	**10,313**	575	1	169	11,266	91.5	953	8.5
	S3			69	**936**	129		1,134	82.5	198	17.5
	S4	2		7	59	**445**		513	86.7	68	13.3
	REM	54	136	231		1	**3,518**	3,940	89.3	422	10.7
	TOT	2,897	1,039	10,913	1,570	577	3,735	20,731	**88.9**	2,303	**11.1**
	Agr. %	93.03	50.14	94.502	59.62	77.12	94.19				

The reading problems were essentially observed in stage 1 (67.2% agreement) and stage 3 (71.6% agreement). The stage 3 sleep epochs were confused primarily with stage 4. Stages 3 and 4 are defined to represent degrees of slow-wave sleep. For some studies, these two stages have been combined into a single stage referred to as slow-wave sleep (SWS), 95% of which is classified correctly. Principal differences are represented by the values adjacent to the main diagonal.

Depressed group

Table 1B gives the agreement matrix derived from the 20 recordings from depressed patients. The total agreement percentage over 20,080 epochs that were visually analyzed was 85.3% between the two experts, i.e. 2,949 misclassified epochs. The reading problems were essentially observed in stage 1 (56.9% agreement) and stage 3 (67.7% agreement). With stages 3 and 4 combined into a single stage (SWS), agreement was 95%.

Insomniac group

Table 1C gives the agreement matrix observed on the 20 benzodiazepine-treated patients with insomnia recordings. The total agreement percentage over 20,731 epochs that were visually analyzed was 88.9% between the two experts, i.e. 2,303 misclassified epochs. The reading problems were essentially observed in stage 1 (60.2% agreement). The SWS agreement percentage was 95%.

Automatic/expert agreement

Similarly, we compared results between the automatic analysis and an expert. The 60 recordings were computed in real time by the neural network. For each recording, we compared the automatic classification and the visual classification of expert 1. Results of comparison between expert 2 and automatic analysis did not differ statistically from those computed with expert 1. The Spec.% row indicates the specificity of the system, i.e. the agreement percentage per stage between expert 1 and automatic analysis with automatic analysis as reference.

Control group

Table 2A gives the agreement matrix derived from the 20 control recordings scored by automatic system and expert 1. The total agreement percentage over 21,138 epochs that were visual analyzed was 84.5%, i.e. 3,272 misclassified epochs. Problems were essentially observed in stage 1 (21.9% agreement) and stage 3 (49.6% agreement). SWS agreement was 90%.

Depressed group

Table 2B gives the agreement matrix derived from the 20 recordings from depressed patients scored by automatic system and expert 1. The total agreement percentage over 20,080 epochs that were visually analyzed was 81.5%, i.e. 3,707 misclassified epochs. Problems were essentially observed in stage 1 (21.6% agreement) and stage 3 (46.7% agreement). SWS agreement was 84%.

Insomniac group

Table 2C gives the agreement matrix derived from the 20 insomniac recordings scored by automatic system and expert 1. The total agreement percentage over 20,731 epochs that were visually analyzed was 81.0%, i.e. 3,936 misclassified epochs. Problems were essentially observed in stage 1 (20.6% agreement) and stage 4 (56.7% agreement). SWS agreement was 86.8%.

Sleep parameter comparison

Table 3 shows the results of sleep parameter comparison among expert 1 (EXP 1), expert 2 (EXP 2) and the automatic system (AUTO). No significant differences between the two experts were observed for the main sleep parameters in the control group, except for the duration of stage 1. In the group of depressed patients, stage 1, stage 3 and SWS duration showed significant variability between readers. In the insomniac group, stage 1 duration and stage 4 latency showed significant variability between readers.

In comparison with expert 1, the automatic sleep stage system showed the same weakness observed in the epoch by epoch comparison. For the control group, the poor detection of stage 1 influenced duration and latency of stage 1. The mixing of stages 3 and 4 induced significant differences in the duration of stages 3 and 4. In the group with depressed patients the duration and latency of stage 1 and the duration of stage 3 were different. In the insomniac group, the duration and latency of stage 1, duration of stage 3 and stage 4 latency differed significantly. In summary, deficiencies observed in epoch by epoch comparison were also observed in the sleep parameter comparison. Overall, differences concerned mainly the detection of stage 1 and the failure to distinguish between stages 3 and 4.

TABLE 2. *Agreement matrix automatic/expert for the three groups*

A. Control		Automatic analysis									
		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
	W	**1,476**	16	186		5	78	1,761	83.8	285	16.2
	S1	212	**142**	166			129	649	21.9	507	78.1
	S2	247	111	**8,223**	119	95	598	9,393	87.5	1,170	12.5
Expert 1	S3	13		283	**632**	331	14	1,273	49.6	641	50.4
	S4	43		56	55	**2,742**	7	2,903	94.5	161	5.5
	REM	133		371		4	**4,651**	5,159	90.2	508	9.8
	TOT	2,124	269	9,285	806	3,177	5,477	21,138	**84.5**	3,272	**15.5**
	Spec. %	69.49	52.79	88.56	78.41	86.31	84.92				

B. Depressed		Automatic analysis									
		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
	W	**3,429**	103	180			143	3,855	88.9	426	11.1
	S1	399	**224**	249	2		162	1,036	21.6	812	78.4
	S2	410	144	**7,576**	214	63	485	8,892	85.2	1,316	14.8
Expert 1	S3	5	1	267	**474**	268	1	1,016	46.7	542	53.3
	S4	3		11	74	**727**		815	89.2	88	10.8
	REM	175	6	342			**3,943**	4,466	88.3	523	11.7
	TOT	**4,421**	478	8,625	764	1,058	4,734	20,080	**81.5**	3,707	**18.5**
	Spec. %	77.56	46.86	87.84	62.04	68.71	83.29				

C. Insomniac		Automatic analysis									
		W	S1	S2	S3	S4	REM	TOT	Agr. %	Diff. n	Diff. %
	W	**2,645**	37	239	1		91	3,013	87.8	368	12.2
	S1	335	**178**	247			105	865	20.6	687	79.4
	S2	455	69	**9,952**	326	41	423	11,266	88.3	1,314	11.7
Expert 1	S3	6		202	**886**	38	2	1,134	78.1	248	21.9
	S4	5		2	215	**291**		513	56.7	222	43.3
	REM	449	64	577	1	6	**2,843**	3,940	72.2	1,097	27.8
	TOT	3,895	348	11,219	1,429	376	3,464	20,731	**81.0**	3,936	**19.0**
	Spec. %	67.91	51.15	88.71	62.00	77.39	82.07				

TABLE 3. *Sleep parameter comparison for the three groups*

	Control								Depressed								Insomniac							
	EXP 1		EXP 2		AUTO		ANOVA		EXP 1		EXP 2		AUTO		ANOVA		EXP 1		EXP 2		AUTO		ANOVA	
	m	σ	m	σ	m	σ	1/2	1/A	m	σ	m	σ	m	σ	1/2	1/A	m	σ	m	σ	m	σ	1/2	1/A
Sleep latency	16	10	18	10	15	10			29	28	27	27	32	36			28	23	27	23	20	17		
Time spent asleep	485	52	489	51	469	52			406	86	421	68	386	94			443	52	447	52	418	52		
Efficiency	92	3	93	3	91	3			81	15	85	11	77	17			86	9	86	9	81	9		
Stage shifts	83	18	95	21	92	19			101	28	94	19	99	25			81	23	80	22	95	23		
NREM time	356	43	361	45	334	47			294	58	312	48	271	73			345	59	353	56	334	60		
REM time	129	26	128	25	135	21			112	44	109	39	116	47			99	33	93	31	84	42		
Awake (minutes)	44	18	39	19	60	32			96	82	79	56	116	93			75	51	72	53	100	54		
Stage 1 (minutes)	16	8	29	17	2	3	*	*	26	13	30	13	9	16	*	*	22	11	26	17	6	8		*
Stage 2 (minutes)	235	44	222	39	238	41			222	59	212	51	219	72			282	53	273	50	283	50		
Stage 3 (minutes)	32	10	37	14	9	9		*	25	12	42	19	15	14	*	*	28	23	39	26	38	38	*	*
Stage 4 (minutes)	73	19	71	19	85	18		*	20	26	27	26	26	28			13	25	14	25	7	20		
SWS (minutes)	104	14	109	16	94	18			46	19	69	23	42	21	*		41	24	54	25	45	29		
Rem sleep latency	67	13	65	12	65	32			64	39	63	40	62	44			96	63	95	62	92	61		
Cycle number	5	1	5	1	5	1			4	1	5	1	4	1			4	1	4	1	4	2		
Stage 1 latency	13	9	12	8	183	229		*	25	27	21	25	64	82		*	28	28	17	14	122	146		*
Stage 2 latency	16	10	18	10	16	11			29	28	27	27	33	35			28	23	27	23	24	19		
Stage 3 latency	29	11	29	11	43	43		*	52	36	52	38	58	44			56	34	63	43	66	61		
Stage 4 latency	35	11	35	12	34	11			84	61	69	46	77	60			70	53	116	52	104	49	*	*

* Significant Holm α-adjustment procedure.

Comparisons between expert 1 (EXP 1), expert 2 (EXP 2) and the automatic system (AUTO). Mean (m) and standard deviation (σ) of each sleep parameter were computed for the 20 nights for each group. Statistical analysis was carried out using analysis of variance for repeated measures. If the "expert" factor (expert 1, expert 2 or automatic analysis) was significant, a Holm α-adjustment procedure (27) was used to characterize the differences. Significant differences between the two experts were marked by an asterisk in the 1/2 column. Significant differences between expert 1 and automatic system were marked by an asterisk in the 1/A column.

Supervision

The above comparisons are between automatic and expert analysis without accounting for uncertainty or unknown states. The major part of the 5% difference observed between experts and automatic system involves these issues.

Across the whole set of 60 nights, the computation of the uncertainty index showed that approximately 19% of the epochs were detected as ambiguous, and half of them were misclassified by automatic scoring. By adding expert supervision for these ambiguous epochs, we were able to correct about 95 epochs per night (9.5% of the recording). Consequently, the automatic/expert agreement rose from 82% to 89%.

Detecting unknowns allows us to reject epochs that differ from the learning set. Across all records, 19.6% of epochs were rejected, and a third of these were misclassified by the automatic system. Expert supervision for unknown epochs improved an average 30 epochs per night (3% of the recording). The automatic/expert agreement rose from 82% to 84%.

DISCUSSION

Our results show both inter-scorer and computer vs. human scorer variability. Inter-reader variability could result from 1) different application of standardized rules of scoring by different readers and 2) a real difficulty in scoring some sleep recordings with the standardized rules (19). The standardized system was developed on and for recordings made for normal young adult volunteers and therefore may be more difficult to apply to abnormal data.

We found that inter-scorer agreement averaged 87.5%, on the basis of 30-second epochs. The population set had little influence on the overall accuracy. Stage 1 was the most variable stage in all groups, probably because it is a short-lasting stage that appears between wake and other sleep stages (principally stage 2 and REM), or after body movement during sleep. REM stage agreement was slightly lower in benzodiazepine-treated insomniac recordings due to beta intrusion and the diminution of eye movements. SWS scoring presented the classical problem of transitions between stages 3 and 4. When stages 3 and 4 were combined into a single stage, SWS, the agreement percentage of SWS was 95%.

Automatic scoring compared to human scoring indicated an average agreement rate of 82.3%. The main differences were observed in stage 1 and the confusion between stages 3 and 4.

Stage 1 scoring was the most unstable stage for automatic scoring in the three groups (20% agreement) because it presented two major characteristics: 1) the a priori probability of stage 1 during the night was very low (between 3% and 6% of the total number of epochs). The classifier used (neural network) converges to the optimum classifier of Bayes that weights the decision by the a priori probability of each class. That means that a small class will be poorly detected in comparison with a large one. Therefore, detection of stage 1 by this automatic detector was unsatisfactory. 2) Stage 1 is a short-lasting stage. Data averaging over the 30-second base time for classification did not allow good temporal resolution for stage 1 detection. Body movement followed by stage 1 might be scored as awake because the movement has a high energy level.

The 82.3% epoch by epoch reliability for our automatic sleep scoring system is comparable with other studies. Moreover, supervision of the ambiguous and unknown epoch automatic decisions offered by the system improved its reliability from 82.3% to 90% after supervision of the recording for only 20% of the night.

The automatic method could be used to quantify other aspects of the sleep process, such as spectral analysis or sleep depth index. However, our purpose in the present study was to validate the system against classical stages of sleep, as defined by the Ad Hoc Committee (47), that are widely used to evaluate sleep structure. It is clear that our automatic technique is less accurate when applied to fragmented sleep with many sleep stage changes or awakenings due to sleep apnea, dementia, alpha-delta sleep or another pathology. This is partially due to the technique being mainly based on continuous EEG activity rather than on phasic events. This method would profit from a combination of techniques that would simultaneously explore sleep staging and the microstructure of phasic events. Further developments in this area are in progress.

Acknowledgements: We thank the sleep team for their support. Thanks should also go to the two experts, C. Ziegler and H. Kempf, for their patience and understanding. Ms G. Savarimouttou-Toussaint, an English teacher, and Dr. P. Bailey are to be thanked for their help with the English version of this paper. Dr. T. Dago is to be thanked for his statistical advice. We also thank Dr. Max Hirshkowitz for the scientific editing of our manuscript.

REFERENCES

1. Frost JD. An automatic sleep analyser. *Electroencephalogr Clin Neurophysiol* 1970;29:88–92.
2. Gaillard JM, Tissot R. Principles of automatic analysis of sleep records with hybrid system. *Comput Biomed Res* 1973;6:1–13.
3. Smith JR, Karacan I, Yang M. Automated analysis of the human sleep EEG. *Waking Sleeping* 1978;2:75–82.
4. Martin WB, Johnson LC, Viglione SS, Naitoh P, Joseph RD, Moses JD. Pattern recognition of EEG–EOG as a technique for all-night sleep stage scoring. *Electroencephalogr Clin Neurophysiol* 1972;32:417–27.

5. Mathieu M, Tirsch W, Pöppl S. Multichannel on line EEG analysis by means of an AR model, with application. In: *Proceedings of the 2nd symposium of the study group for EEG-methodology.* GK Schenk Jongny/Vevey, May 1975, 475–85.
6. Gath I, Bar-on E. Computerized method for scoring of polygraphic sleep recordings. *Comp Prog Biomed* 1980;11:217–23.
7. Stanus E, Lacroix B, Kerkhofs M, Mendlewicz J. Automated sleep scoring: a comparative reliability study of two algorithms. *Electroencephalogr Clin Neurophysiol* 1987;66:448–56.
8. Kuwahara H, Higashi H, Mizuki Y, Matsunari S, Tanaka M, Inagana K. Automatic real-time analysis of human sleep stages by an interval histogram method. *Electroencephologr Clin Neurophysiol* 1988;70:220–9.
9. Ray SR, Lee WD, Morgan CD, Airth-Kindree W. Computer sleep stage scoring: an expert system approach. *Int J Biomed Comput* 1986;19:43–61.
10. Kubat M, Pfurtscheller G, Flotzinger D. AI-based approach to automatic sleep classification. *Biol Cybern* 1994;5:443–8.
11. Roberts S, Tarassenko L. Analysis of sleep EEG using a multilayer network with spatial organisation. *IEEE proceedings, part F, radar signal processing, special issue on artificial neural networks,* vol. 139. 1992;6:420–5.
12. Schaltenbrand N, Lengelle R, Macher JP. Neural network model: application to automatic analysis of human sleep. *Comput Biomed Res* 1993;26:157–71.
13. Hasan J. Differentiation of normal and disturbed sleep by automatic analysis. *Acta Physiol Scand* 1983;Suppl 526:1–103.
14. Kubicki St, Höller L, Berg I, Pastelak-Price C, Dorow R. Sleep EEG evaluation: a comparison of results obtained by visual scoring and automatic analysis with the Oxford sleep stager. *Sleep* 1989;12:140–9.
15. Ferri R, Ferri P, Colognola M, Petrella MA, Musumeci SA, Bergonzi P. Comparison between the results of an automatic and visual scoring of sleep EEG recordings. *Sleep* 1989;12:354–62.
16. Kim Y, Kurachi M, Horita M, Matsuura K, Kamikawa Y. Agreement in visual scoring of sleep stages among laboratories in Japan. *J Sleep Res* 1992;1:58–60.
17. Rechtschaffen A, Kales A, Eds. *A manual of standardized terminology techniques and scoring system for sleep stages of human subjects.* Washington DC: Public Health Service, U.S. Government Printing Office, 1968.
18. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders, third edition revised.* Washington DC: American Psychiatric Association, 1987.
19. Hamilton H. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967;16:39–48.
20. Borbély AA, Akerstedt T, Benoit O, Holsboer F, Oswald P. Hypnotics and sleep physiology: a consensus report. *Eur Arch Psychiatry Clin Neurosci* 1991;241:13–21.
21. Lengelle R, Hao Y, Schaltenbrand N, Denoeux T. Ambiguity and distance rejection using multilayer neural networks. In: Dagli CH, Kumara SRT, Shin YC, eds. *Intelligent engineering through artificial neural networks.* ASME Press, 1991:299–304.
22. Rumelhart DE, McClelland JL. *Parallel distributed processing: explorations in the microstructure of the cognition.* Boston, MA: MIT Press, 1986.
23. Kubicki St, Herrmann WM, Höller L. Critical comments on the rules by Rechtschaffen and Kales concerning the visual evaluation of EEG sleep records. In: Kubicki St, Herrmann WM, eds. *Methods of sleep research.* Stuttgart: Gustav Fisher, 1985:19–35.
24. Carpenter GA, Grossberg S. ART2: self-organization of stable category recognition codes for analog input patterns. *Appl Optics* 1987;26:4919–30.
25. Yin H, Lengelle R, Gaillard P. NeoART: une Variante du Réseau ART2 pour la Classification. *Proceedings of the third international workshop on neural networks and their applications* (Neuro-Nîmes 1990). Nîmes, France: November 12–16, 1990: 171–179.
26. Lengelle R, Schaltenbrand N, Cornu Ph, Gaillard P. Pattern recognition, using neural networks: comparison to the nearest neighbour rule. *Proceedings of the 1st IFAC symposium AIPAC '89.* Husson R, ed. Nancy, France: July 3–5, 1989;2:97–102.
27. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist* 1979;6:65–70.